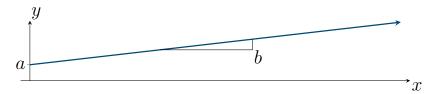
# 4.3: Modeling Linear Trends

### Definition.

The **regression line** is a model used for making predictions about *future* observed values. The equation of the regression line is

$$y = a + bx$$

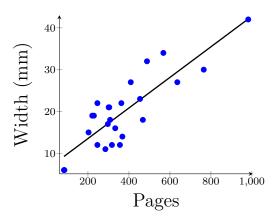
where a is the y-intercept and b is the slope.



- $\bullet$  The input variable x is also know as the
  - Independent variable
  - Predictor variable
  - Explanatory variable
- ullet The output variable y is known as the
  - Dependent variable
  - Predicted variable
  - Response variable

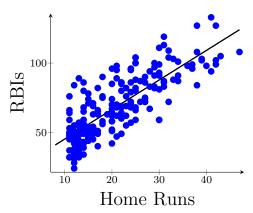
**Example.** Below is a scatterplot comparing number of pages a book has against the width of the book. Interpret the intercept and the slope of the regression line.

Predicted Width=6.22+0.0366 Pages



**Example.** Below is a scatterplot comparing the number of home runs and RBIs in the 2016 season. Interpret the intercept and slope of the regression line.

Predicted RBI= $23.84+2.13\,\mathrm{HR}$ 



## Definition.

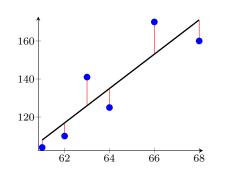
Now we define the formula of the regression line:

$$y = a + bx$$

Where

$$b = r \frac{s_y}{s_x}$$
 and  $a = \overline{y} - b\overline{x}$ .

These formulae minimize the residual error: Try this!



**Example.** Below are the heights and weights of six women:

Heights	61	62	63	64	66	68
Weights	104	110	141	125	170	160

From this we get

$$\overline{x} = 64$$

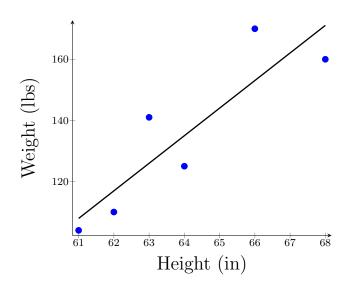
$$\overline{y} = 135$$

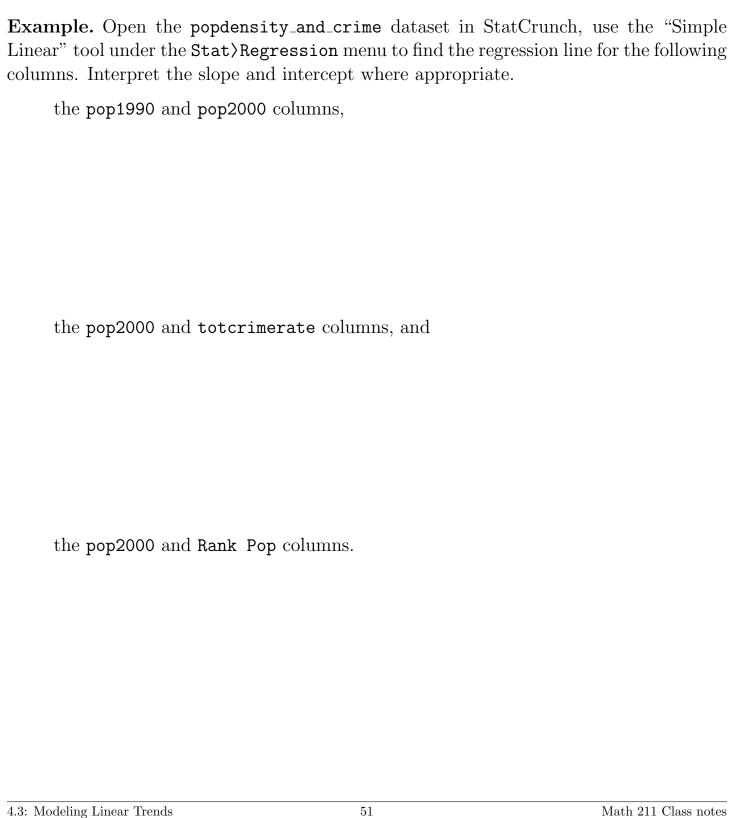
$$r = 0.881$$

$$s_x = 2.608$$

$$s_y = 26.728$$

Find the equation of the regression line.





## 4.4: Evaluating the Linear Model

#### Guidelines:

- Don't fit linear models to nonlinear associations!
- Correlation is not causation
- Beware of outliers (a.k.a. influential points)
- Don't extrapolate (make predictions beyond the range of the data)

### Definition.

The **coefficient of determination** is the correlation coefficient coefficient squared:

 $r^2$ 

This is sometimes also called r-squared.